Paper 1

Title: Power Calculations for Binary Moderator in Cluster Randomized Trials

Authors and Affiliations:

Jessaca Spybrook
Western Michigan University

Ben Kelcey
University of Cincinnati

Background / Context:
Cluster randomized trials (CRTs), or studies in which intact groups of individuals are randomly assigned to a condition, are becoming more common in the evaluation of educational programs, policies, and practices. For example, a search on the website for the Institute of Education Sciences (IES) suggests that the National Center for Educational Research (NCER) funded around 7 randomized trials in 2004. The same search for 2011 yielded around 36 funded randomized trials, or approximately 5 times as many trials. The website for the National Center for Education Evaluation and Regional Assistance (NCEE) reveals they have launched over 30 evaluation studies in the past decade, the majority of them utilizing a randomized trial. Clearly there are a large number of randomized trials of educational programs, policies, and practices either complete or currently in the field.

The overarching goal of these randomized trials is generate rigorous evidence of whether or not a program works. In statistical terms, this is often referred to as the main effect of treatment. In the past 15 years, the field has made substantial progress in terms of how to design CRTs and how to calculate the statistical power for the main effect of treatment. However, designing a study to detect the main effect of treatment may not be sufficient. It is quite reasonable that *context matters* in these studies and thus designing studies to examine *for whom and under what conditions a program is effective* is critical. In order to do this, studies must also be designed to detect moderator effects. The power to detect moderator effects at the student, cluster, or site level in CRTs has received much less attention in the literature than the power for the main effect of treatment. Some of the recent work includes: Bloom (2005) provides power calculations for individual and cluster level moderators in a 2-level CRT; Raudenbush and Liu (2000) show power calculations for site-level moderator effects for multisite trials in which individuals are randomly assigned within sites; Hedges and Pigott (2004) discuss power calculations for moderator effects in the context of a meta-analysis which has direct comparisons to multisite cluster randomized trials; and Spybrook (in press) examines power for individual and cluster level moderators for CRTs but not site level moderators.

Purpose / Objective / Research Question / Focus of Study:
The purpose of this paper is to extend the work on power calculations for moderator effects to include moderator effects at any level for the following 4 types of CRTs: 2-level CRT, 3-level CRT, 3-level multisite cluster randomized trial (MSCRT), and 4-level MSCRT. In addition to providing the calculations and R code to do the calculations, we start to develop intuition around the minimum detectable effect size for moderator effects using sample sizes from CRTs in the field of education.

Significance / Novelty of study:
This paper represents the next step towards building the capacity of researchers to design CRTs that move beyond the main effect of treatment. The three primary power programs for CRTs, *Optimal Design Plus* (Raudenbush, Spybrook, Congdon, Liu, Martinez, Bloom, & Hill, 2011), *CRT Power* (Borenstein & Hedges, 2011), and *Power UP* (Dong & Maynard, 2013) do not routinely allow users to calculate power for moderator effects. The calculations and R code in this paper provide an accessible resource for researchers calculating power for moderator effects.

Statistical, Measurement, or Econometric Model:
Given the space limitations in this proposal, we provide the models for the 3-level MSCRT and the power calculations for the main effect of treatment and a site level moderator. Details for all of the designs will be provided in the full paper.
Main effect of Treatment

We begin with the power for main effect for treatment, since this is typically the primary effect of interest and provides intuition for the calculations for moderator effects. Suppose we have a study in which schools are the unit of randomization, but they are blocked by district. That is, within each district, schools are randomly assigned to condition and students are nested within schools, a 3-level MSCRT. In the case of no moderator, the student level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2) \qquad [1]$$

for $i \in \{1,2,...,n\}$ persons per cluster, $j \in \{1,2,...,J\}$ clusters and $k \in \{1,2,...,K\}$ sites,

where π_{0jk} is the mean for cluster *j* in site *k*; e_{ijk} is the error associated with each person; and σ^2 is the within-cluster variance.

The level-2 model, or cluster-level model, is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} W_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi) \qquad [2]$$

where β_{00k} is the mean for site *k*; β_{01k} is the treatment effect at site *k*; W_{jk} is a treatment contrast indicator, ½ for treatment and -½ for the control; r_{0jk} is the random effect associated with each cluster; and τ_π is the variance between clusters within sites.

The level-3 model, or site-level model, is:

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad \text{var}\ (u_{00k}) \sim \tau_{\beta_{00}}$$

$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad \text{var}\ (u_{01k}) \sim \tau_{\beta_{11}} \qquad \text{cov}(u_{00k}, u_{01k}) = \tau_{\beta_{01}} \qquad [3]$$

where γ_{000} is the grand mean; γ_{010} is the average treatment effect ("main effect of treatment");

u_{00k} is the random effect associated with each site mean; u_{01k} is the random effect associated with each site treatment effect; $\tau_{\beta_{00}}$ is the variance between site means; $\tau_{\beta_{11}}$ is the variance between sites on the treatment effect; and $\tau_{\beta_{01}}$ is the covariance between site-specific means and site-specific treatment effects. Note that we allow the treatment effect to vary randomly across sites, however, we could also treat this as a fixed effect.

The estimate of the treatment effect and the variance of the estimated treatment effect are::

$$\hat{\gamma}_{010} = \bar{Y}_E - \bar{Y}_C$$

$$Var\left(\hat{\gamma}_{010}\right) = \left[\tau_{\beta_{11}} + 4\left(\tau_{\pi} + \sigma^2 / n\right)/ J\right]/ K \qquad [4]$$

The power for the test for the main effect of treatment for the 3-level MSCRT, $H_0 : \gamma_{010} = 0$, follows the same logic as the power for the main effect of treatment for the 2-level CRT (Raudenbush, 1997). The *F*statistic in this case though is a ratio $MS_{treatment}$ to the $MS_{treatmentbycluster}$. The ratio of expected mean squares is equivalent to 1+ λ, where the noncentrality parameter is defined as:

$$\lambda = \frac{\gamma_{010}^2}{\left[\tau_{\beta_{11}} + 4\left(\tau_{\pi} + \sigma^2 / n\right)/ J\right]/ K} \quad \text{or} \quad \lambda = \frac{\delta^2}{\left[\sigma_{\delta}^2 + 4\left(\rho + (1-\rho)/ n\right)/ J\right]/ K} \qquad [5]$$

where $\delta = \frac{\gamma_{010}}{\sqrt{\tau_{\pi} + \sigma^2}}, \rho = \frac{\tau_{\pi}}{\tau_{\pi} + \sigma^2}$, $\sigma_{\delta}^2 = \frac{\tau_{\beta_{11}}}{\tau_{\pi} + \sigma^2}$

We standardize the parameters by the sum of the within site variance. As the noncentrality parameter increases, the power of the test increases.

Site Moderator Effects

As mentioned above, given the space limitations for the proposal, we provide the models with a site level moderator. Level-1 and level 2 remain the same as equations 1 and 2. However, in level 3 we assume there are an equal number of rural and urban sites (site type).

The level-3 model, or site-level model, is:

$$\beta_{00k} = \gamma_{000} + \gamma_{001} S_k + u_{00k} \qquad \text{var}\ (u_{00k}) \sim \tau_{\beta_{00|s}}$$

$$\beta_{01k} = \gamma_{010} + \gamma_{011} S_k + u_{01k} \qquad \text{var}\ (u_{01k}) \sim \tau_{\beta_{11|s}} \qquad [6]$$

where γ_{000} is the grand mean; γ_{001} is the effect of site type on the mean; γ_{010} is the average treatment effect; γ_{011} is the site type (urban or rural) by treatment interaction; S_k is site contrast indicator, ½ for rural and -½ for urban; u_{00k} is the random effect associated with each site mean; u_{01k} is the random effect associated with each site treatment effect; $\tau_{\beta_{00}}$ is the residual variance between site means; $\tau_{\beta_{11}}$ is the residual variance between sites on the treatment effect; and $\tau_{\beta_{01}}$ is the residual covariance between site-specific means and site-specific treatment effects. Note that we allow the treatment effect to vary randomly across sites, however, we could also treat this as a fixed effect.

The power for the test of the site-level moderator for the 3-level MSCRT, $H_0 : \gamma_{011} = 0$, follows the same logic as the power for the main effect of treatment. However the noncentrality parameter in this case is:

$$\lambda = \frac{{\delta_s}^2}{\left[4\sigma_{\delta|s}^2 + 16\left(\rho + (1-\rho)/ n\right)/ J\right]/ K} \qquad [7]$$

Note that δ_s is simply the standardized site moderator effect and $\sigma^2_{d|s}$ is the residual variance in the treatment effect across sites. A quick comparison of the noncentrality parameters associated with the test for the main effect of treatment (equation 5) and the site moderator effect (equation 7) reveal the following. The site level moderator does reduce the site by treatment variance, $\sigma^2_{d|s}$, however, the within site variance, $\left[\rho + (1-\rho)/n\right]/J$ is 4 times larger for the site moderator effect than in the case for the main effect of treatment. This suggests much larger sample sizes are needed to detect site moderator effects.

Findings / Results:
Larger sample sizes are required order to detect a site level moderator compared to the main effect of treatment. Although not shown in this proposal, this finding also holds for a cluster level moderator. However, for CRTs, the power to detect an individual level moderator may be larger than the main effect of treatment because the number of individuals is the most influential sample size for detecting individual level moderator effects (Bloom, 2005; Spybrook, in press).

Conclusions:
Designing studies to detect not only whether or not an intervention works, but for whom or under what circumstances is critical. The results from this study suggest that in many cases, if a study is powered to detect a reasonable main effect of treatment and it has a reasonable number of individuals per cluster, then it will also be powered to detect an individual level moderator (although not shown in this proposal). However, the sample size requirements for adequate power to detect a cluster level moderator or site level moderator exceed the sample size requirements to power a study to detect a main effect of treatment. This presents an important challenge to researchers designing CRTs and to funding agencies supporting CRTs. For the researchers, it is critical to perform these calculations so that they are aware of whether or not they are powered to detect moderator effects. For funders, this suggests that given current levels of funding, studies may not be able to detect more than the main effect of treatment.

Appendix A. References

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments evolving analytic approaches* (pp. 115-172). New York: Russell Sage.

Borenstein, M., & Hedges, L.V. (2011). *CRT-Power*. Biostat, Inc.

Hedges, L.V., & Pigott,, T.D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4), 426-445.

Dong, **N.,** & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S.W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). *Optimal Design Plus Empirical Evidence* (Version 3.0)

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

Spybrook, J. (in press). Detecting intervention effects across context: An Examination of the Power of Cluster Randomized Trials. *Journal of Experimental Education*.